Mining Quantitative Association Rules in Protein Sequences

Nitin Gupta[1], Nitin Mangal[2], Kamal Tiwari, and Pabitra Mitra

[1] Bioinformatics Group, Dept. of Computer Science,
University of California, San Diego,
3859 Miramar Street #D, La Jolla, CA 92037, USA
nitiniitk@yahoo.com
[2] Department of Computer Science and Engineering,
Indian Institute of Technology Kanpur - 208016, India
mangal_iitk@yahoo.com, pmitra@cse.iitk.ac.in

Abstract. Lot of research has gone into understanding the composition and nature of proteins, still many things remain to be understood satisfactorily. It is now generally believed that amino acid sequences of proteins are not random, and thus the patterns of amino acids that we observe in the protein sequences are also non-random. In this study, we have attempted to decipher the nature of associations between different amino acids that are present in a protein. This very basic analysis provides insights into the co-occurrence of certain amino acids in a protein. Such association rules are desirable for enhancing our understanding of protein composition and hold the potential to give clues regarding the global interactions amongst some particular sets of amino acids occuring in proteins. Presence of strong non-trivial associations suggests further evidence for non-randomness of protein sequences. Knowledge of these rules or constraints is highly desirable for the in-vitro synthesis of artificial proteins.

Keywords: Data mining, quantitative association rule mining, protein composition.

1 Introduction

Proteins are important constituents of cellular machinery of any organism. Recombinant DNA technologies have provided tools for the rapid determination of DNA sequences and, by inference, the amino acid sequences of proteins from structural genes [1]. The proteins are sequences made up of 20 types of amino acids. Each amino acid is represented by a single letter alphabet, see Table 1. Each protein adopts a unique 3-dimensional structure, which is decided completely by its amino-acid sequence. A slight change in the sequence might completely change the functioning of the protein.

The heavy dependence of protein functioning on its amino acid sequence has been a subject of great anxiety. Research has been done to determine the information content per amino acid in proteins by Yockey [2] and Strait & Dewey [3].

G.J. Williams and S.J. Simoff (Eds.): Data Mining, LNAI 3755, pp. 273–281, 2006.

Table 1. Single letter codes of amino acids

S.No.	AA Code	Full-Name
1	A	Alanine
2	C	Cysteine
3	D	Aspartic Acid
4	E	Glutamic Acid
5	F	Phenylalanine
6	G	Glycine
7	H	Histidine
8	I	Isoleucine
9	K	Lysine
10	L	Leucine
11	M	Methionine
12	N	Asparagine
13	P	Proline
14	Q	Glutamine
15	R	Arginine
16	S	Serine
17	T	Threonine
18	V	Valine
19	W	Tryptophan
20	Y	Tyrosine

There has been a continuing debate on whether the amino acid seqeunces of proteins are random or have statistically significant deviations from random sequences. White & Jacobs [5] have shown that any sequence chosen randomly from a large collection of nonhomologous proteins has a 90% or better chance of having a lengthwise distribution of amino acids that is indistinguishable from the random expectation regardless of amino acid type. They claimed that proteins have evolved from random sequences but have developed significant deviations from randomness during the process of evolution. Pande et al [4] mapped protein sequences to random walks to detect differences in the trajectories of a Brownian particle. They found pronounced deviations from pure randomness which seem to be directed towards the minimization of energy in the 3D structure.

In this study, we take a further step in this direction by trying to predict if there are any co-occurrence patterns among the 20 amino-acids. We have attempted to find out rules that can tell that occurrence of one amino-acid is more likely when another amino-acid is present or absent. Such rules are called "association rules", and the corresponding technique is called "association rule mining" (ARM). In ARM terminology, the amino- acids may be considered as items, and the protein sequences as "baskets" containing items. See the next section for a introduction to association rule mining. Proteins are polymers of length usually in hundreds. Since the length is much larger, all the 20 amino acids are present in majority of proteins, and thus we will not be able de-

duce any significant rule just based on presence or absence. To obtain more meaningful association rules in this context, we have incorporated the normalized frequencies of amino-acids observed in each protein, and also discovered "quantitative association rules", which tell that if one amino-acid A is present with a f_1 frequency, another amino-acid B is likely to be present with f_2 frequency. Our quantitative association rule mining procedure [8] enables us to find these numbers f_1 and f_2.

The organization of this paper is as follows; the next section gives an overview of association rule mining. Section 3 describes how we have implemented association rule mining for finding quantitative rules in proteins. Section 4 shows the rules that we have obtained. The next section discusses these results and concludes the outcomes of this study, followed by future work describing how this study can be extended.

2 Association Rule Mining

Before we begin with the description of our algorithm, it will be helpful to review some of the key concepts of association rule mining. We use the same notation as used in [9]. Let $I = \{i_1, ..., i_k\}$ be a set of k elements, called *items*. Let $B = \{b_1, ..., b_n\}$ be a set of n subsets of I. We call each $b_i \subseteq I$ a *basket* of items. For example, in the market basket application, the set I consists of the items stocked by a retail outlet and each basket is the set of purchases from one register transaction. Similarly, in the "document basket" application, the set I contains all dictionary words and proper nouns, while each basket is a single document in the corpus. Note that the concept of a basket does not take into account the ordering or frequency of items that might be present. An association rule is intended to capture a certain type of dependence among items represented in the database B. Specifically, we say that $i_1 \rightarrow i_2$ if the following two hold

1. i_1 and i_2 occur together in at least $s\%$ of the n baskets (the *support*).
2. Of all the baskets containing i_1, atleast $c\%$ also contain i_2 (the *confidence*).

This definition is also extended to $I \rightarrow J$, where I and J are disjoint sets of items instead of single items. Let us consider an example of a document basket application. The baskets in this case are many short stories that are available at our disposal, while the items within each basket are the words. A reader might observe that stories which contain the word "sword" also frequently contain the word "blood". This information can be represented in the form of a rule as:

$$sword \rightarrow blood$$

$$[support = 5\%, confidence = 55\%] \tag{1}$$

Rule support and confidence are the two measures of rule interestingness [10]. They respectively reflect the usefulness and certainty of discovered rules. A support of 5% for an association rule means that 5% of stories under analysis

show that "blood" and "sword" occur together. A confidence of 55% means that 55% of the stories that contain the word "sword" also contain the word "blood". Typically, associations rules are considered interesting if they satisfy both minimum support threshold and a minimum confidence threshold. Such threshold can be set by users or domain experts. As pointed out in [9], it should be noted that the symbol $\rightarrow$ is misleading since such a rule does not correspond to real implications; clearly, the confidence measure is merely an estimate of the conditional probability of i_2 given i_1.

2.1 The Apriori Algorithm

The most commonly used approach for finding association rules is based on the Apriori algorithm [6]. Apriori employs an iterative approach known as a level-wise search, where k-itemsets (sets containing k items) are used to explore $(k+1)$-itemsets. First, the set of frequent (i.e. having more than the minimum support) 1-itemsets is found. This set is used to find set of frequent 2-itemsets, which is used to find the set of frequent 3-itemsets, and so on, until no more frequent k-itemsets can be found. The efficiency of the level-wise generation of frequent itemsets is improved by using the Apriori property which says that all nonempty subsets of a frequent itemset must also be frequent. This is easy to observe, because if an itemset I does not satisfy the minimum support threshold, then the set $I' = I \vee \{i_{new}\}$, containing all elements of I and an extra element i_{new}, cannot occur more frequently than I, and thus cannot satisfy the minimum support threshold.

2.2 Quantitative Association Rules

While the association rule model described above suffices for many applications, it is not adequate when the frequency of each item in the basket is variable and cannot be ignored. For example, in the previously considered example, a user might be interested in the rules of the form:

$$sword_{30-35} \wedge war_{14-16} \rightarrow blood_{50-52} \tag{2}$$

This rule represents that a story that contains between 30 to 35 occurrences of "sword" and 14 to 16 occurrences of "war", is also likely to contain 50 to 52 references of "blood". Such rules are called quantitative association rules.

The ARCS system [11] for mining quantitative association rules is based on rule clustering. Essentially this approach maps pairs of quantitative attributes onto a multi-dimensional grid, with the number of dimensions equaling the number of quantitative attributes considered. The grid is then searched for clusters of points, from which the association rules are generated. Techniques for mining quantitative rules based on x-monotone and rectilinear regions were presented in [7]. Approach proposed in [8] works by fine-partitioning the values of the quantitative attributes, and then generating rules of interest.

3 Algorithm

Our implementation is based on the partitioning approach described in [8]. We consider 20 attributes in proteins, each related to an amino acid. The value of each attribute in a basket (here protein) is the frequency of the corresponding amino acid in the protein. Since the proteins are of varying lengths, we normalize this frequency by dividing by the length of the protein.

The main steps in the algorithm are as follows:

1. Partition the attributes: We have divided each of the 20 attributes into 10 intervals. In [8], the authors have discussed the notion of partial completeness to quantify the amount of information lost due to partitioning. It has been further shown that for a given number of partitions, equi-depth partitioning (each partition having equal support) gives the minimum loss of information, and is thus optimal. Thus, we have used equi-depth partitioning in our method. For the sake of completeness and comparison, we have also experimented with equi-distant partitioning, in which all intervals are of equal length.
2. The intervals/partitions are mapped into consecutive integers, which are used to represent the intervals. The order of intervals is preserved in the mapping.
3. Find the support for each of the intervals. Also the consecutive intervals are combined as long as their support is less than a predetermined maximum support. This is actually needed in case of equi-distant partitioning when some of the intervals may have very small support and thus it makes sense to combine them with the adjoining intervals. In equi-depth partitioning, all intervals have equal support, and thus this problem does not arise. We identify the set of all intervals which have more than a minimum support *minsup*. This is called the set of *frequent* items.
   Next we find all sets of items whose support is greater than *minsup*. These are called the frequent itemset, and the algorithm is based on the Apriori algorithm, discussed in the previous section.
4. The frequent itemsets are used to generate association rules. each itemset can give rise to number of association rules by dividing into two parts: antecedents and consequences. For example, an itemset {P,Q,R} can lead to the following rules

   - $P \rightarrow Q \wedge R$
   - $Q \wedge R \rightarrow P$
   - $P \wedge Q \rightarrow R$
   - $R \rightarrow P \wedge Q$
   - $Q \rightarrow P \wedge R$
   - $P \wedge R \rightarrow Q$

   The confidence *conf* for each of the rules is determined as the conditional probability of conclusion given precedent. For example, for the rule

$$P \wedge Q \rightarrow R, conf = support\{P, Q, R\}/support\{P, Q\}$$

If the confidence is greater than a pre-determined minimum confidence, *min-conf*, the rule is kept, otherwise it is removed.

4 Results

The protein sequences are taken from the SCOP Astral File v1.63 [12], containing only those sequences which are less than 40% homologous to each other. This reduces the bias in favour of highly populated families as compared to sparse ones. The sequences with length less than 100 or more than 500 are not considered. This gives us a set of 3728 non-homologous amino-acid sequences representing the different types of proteins. In this study our focus is on deriving associations applicable to all proteins in general.

Figure 1 shows the rules obtained with minimum support of 30 proteins. We have obtained 12 association rules, which have confidence more than 50%. The universe of chains that can be built from 20 amino acids is extremely large and diverse. In light of this fact, the confidence and support of the rules presented in Figure 1 are quite significant.

As an example, the eighth rule indicates that proteins containing large amounts of Arginine(R) and very low amount of Serine(S) are likely to contain no Cysteine (C). Cysteines are the amino acids that participate in the formation of disuphide bonds in the amino acids. This rule implies that presence of large amounts of Arginine without compensating Serine will hinder the ability of a protein to form the disulphide bonds. Such rules provide some insight into the interaction and role of these amino acids in proteins, and have important consequences in the emerging field of *synthetic biology* where biological entities are designed and synthesized in the lab.

Rule	Confidence(%)	Support
<G,52..500> ^ <S,45..500> => <E,0..16>	64.7	33
<E,0..16> ^ <L,0..26> => <T,40..500>	60.9	39
<E,0..16> ^ <M,0..2> => <T,40..500>	59.6	31
<L,0..26> ^ <S,45..500> => <T,40..500>	55.0	38
<E,0..16> ^ <L,0..26> => <G,52..500>	54.6	35
<I,0..13> ^ <R,39..500> => <N,0..8>	54.4	43
<K,0..11> ^ <S,45..500> => <E,0..16>	54.2	32
<R,39..500> ^ <S,0..14> => <C,0..0>	53.5	30
<K,0..11> ^ <N,0..8> => <R,39..500>	53.4	31
<P,35..500> ^ <R,39..500> => <N,0..8>	52.6	30
<L,64..500> ^ <P,35..500> => <N,0..8>	51.7	30
<I,0..13> ^ <N,0..8> => <R,39..500>	50.5	43

Fig. 1. Associations obtained using equi-depth partitioning. Each interval (contained in angular brackets) has an amino acid, and frequency range with protein length scaled to 500. The support is the number of proteins in our dataset of 3728 proteins containing all the intervals present in the association rule.

Rule	Confidence(%)	Support
<I,40..49> ^ <R,20..29> => <W,0..9>	94.5	139
<C,0..9> ^ <F,10..19> ^ <P,10..19> ^ <V,40..49> => <W,0..9>	94.4	102
<C,0..9> ^ <I,40..49> ^ <N,10..19> => <W,0..9>	94.1	112
<I,40..49> ^ <L,40..49> => <W,0..9>	93.6	118
<A,60..79> ^ <P,20..29> ^ <W,0..9> => <C,0..9>	93.6	103
<C,0..9> ^ <H,0..9> ^ <S,20..29> ^ <Y,0..9> => <W,0..9>	93.6	104
<D,20..29> ^ <P,10..19> ^ <V,40..49> => <W,0..9>	93.5	101
<A,60..79> ^ <T,20..29> ^ <W,0..9> => <C,0..9>	93.3	126
<H,0..9> ^ <N,10..19> ^ <Y,0..9> => <W,0..9>	93.2	151
<Q,10..19> ^ <S,20..29> ^ <Y,0..9> => <W,0..9>	93.2	110
<H,0..9> ^ <V,40..49> ^ <Y,0..9> => <W,0..9>	93.1	108
<C,0..9> ^ <S,20..29> ^ <T,20..29> ^ <V,40..49> => <W,0..9>	93.1	109
<C,0..9> ^ <H,0..9> ^ <N,10..19> ^ <Y,0..9> => <W,0..9>	93.0	121

Fig. 2. Associations obtained using equi-distant partitioning. Representation is same as in Figure 1. Note that consequence part in all the rules contains either Cysteine (C) or Tryptophan (W). See results section for the discussion of this behavior.

To see how the performance of the algorithm changes when equi-distant rules are used, we created 10 intervals of equal length with frequency ranges 0-9, 10-19,..., 80-89 and 90-500. Note the last interval has been stretched to accommodate any arbitrarily high frequency, which is extremely rare. The proteins lengths are scaled to 500 and the frequencies are increased or decreased in proportion. The association rules obtained from this approach are shown in Figure 2. As expected, the method gets heavily biased in favour of those intervals which have very high supports. For example, Cysteine(C) and Tryptophan(W) are the less frequent amino acids in proteins; in most proteins the frequency of these amino acids is close to zero. Thus the lowermost intervals for these two amino acids get very high support value, and thus generate association rules with very high support and confidence. Note that these rules, inspite of high confidence and support, are not useful to biologists. The consequence part of these rules say that Cysteine and Tryptophan occur in range 0-9, which is trivially known for majority of the proteins.

5 Discussions and Conclusion

We have used quantitative association rule mining to discuss global associations between amino acids in proteins. We call the associations global because the rules are not forced to be based on contiguous set of amino acids, and thus can capture global correlations as well.

The amino acid frequencies are divided into intervals to build the rules. We observe that equi-depth partitioning gives 12 association rules involving various amino acids. The use of equi-distant partitioning gives skewed results, because the relative frequencies of amino acids in the proteins are highly different and equi-distant partitioning results in some very highly populated and some very sparsely populated partitions. This is in line with the conclusion about supremacy of equi-depth approach drawn from the concept of partial completeness in [8].

An important property of our approach is that it can discover rules based not only the presence of amino acids, but also on absence. For example, the eighth rule in Figure 1 has the consequence which says C is likely to be absent. This is a significant difference from the standard motif based works, which are framed only the basis of presence of an amino acid. We acknowledge the fact that absence of a particular amino acid can also be important in the structure and/or function a protein.

To the best of our knowledge, this is the first systematic study to discover global associations between amino acids. The rules obtained here present the constraints in the composition of proteins, and will prove very important in the design and synthesis of artificial peptides, outside the cell. The pharmaceutical industry is gradually shifting from *small molecule* drugs to *biologics* which are synthetic peptides, and is likely to benefit from the availability of knowledge about the rules governing the composition of peptides found in the nature.

This work can be extended in following ways:

- The rules generated in this study are very interesting, and non-trivial. Experimental verification of these methods is a big challenge, and there is no easy way to do that. One strategy could be to design synethetic amino acid chains that violate the rules obtained here, and study their physico-chemical properties in-vitro to see if they behave differently.
- Our approach has been based on partitions approach proposed in [8]. It is possible to use other approaches as well, and it is to be seen if they result in some more interesting rules.
- Instead of finding rules based on whole set of proteins, specialized rules can be found for different classes of proteins. This, however, requires a larger protein dataset containing sufficient number of distinct and non-homologous represenentatives in each class.

Acknowledgment

We thank Dr. Somenath Biswas (Computer Science, IIT Kanpur) for valuable discussions.

References

1. Branden, C. and Tooze, J. *Introduction to Protein Structure* (Garland Publishing, New York, 1991).
2. Yockey, H. P. (1977). On the information content of cytochrome. *J. Theor. Biol.* 67, 147-151.
3. Strait, B.J.& Dewey, G.(1996). The Shannon information entropy of protein sequences. *Biophys. J.* 71, 148-155.
4. Pande, S. V., Grosberg, A. Y. & Tanaka, T. (1994). Non-randomness in protein sequences: evidence for a physically driven stage of evolution? *Proc. Natl. Acad. Sci. U.S.A.* 91, 12972-12975.

5. White, S. H. & Jacobs, R. E. (1993). The evolution of proteins from random amino acid sequences - I. Evidence of proteins from the lengthwise distribution of amino acids in modern proteins. *J. Mol. Evol.* 36, 79-95.
6. Agrawal, R. and Srikant, R. (1994). Fast algorithms for mining association rules. In *Proc of the 20th Int'l Conference on Very Large Databases*, Santiago, Chile, September'94.
7. Fukuda, T., Morimoto, Y., Morishita, S. and Tokuyama, T. (1996) Data mining using two-dimensional optimized association rules: Scheme, algorithms, and visualization. In *Proc. 1996 ACM-SIGMOD Int. Conf. Management of Data*, pp 13-23, Montreal, Canada.
8. Srikant, R. and Agrawal, R. (1996). Mining quantitative association rules in large relational tables. *Proc. ACM SIGMOD.*
9. Brin, S., Motwani, R., and Silverstein, C. (1997). Beyond market basket: Generalizing association rules to correlations. In *Proc. 1197 ACM SIGMOD*, pp 265-276. Tuscon, AZ.
10. Han, J. and Kamber, M. Data Mining: Concepts and Techniques. *Morgan Kaufmann Publishers*, San Francisco, 2001.
11. Lent, B., Swami, A. and Widom, J. (1997). Clustering association rules. In *Proc. Int'l Conf. Data Engineering (ICDE'97)*, pp220-231, England.
12. http://scop.mrc-lmb.cam.ac.uk/scop/